DIGITAL PRESERVATION STRATEGIES FOR AV CONTENT

Matthew Addis[1], Richard Wright[2], Rajitha Weerakkody[2]

[1]IT Innovation Centre, University of Southampton, UK, [2]BBC Research and Development, London, UK

ABSTRACT

The mass digitisation of analogue archive holdings plus the transition to tapeless production for new content means that AV archives now face the prospect of file-based archiving solutions using IT storage technology. But what is the long-term Total Cost of Ownership (TCO) of these systems, which file formats should be used, what storage technologies make sense, what are the risks involved, what is the additional cost of managing these risks, and what new software approaches can be applied? These issues are being explored by major broadcasters, national archives and technology specialists in the PrestoPrime and AVATAR-m projects.

INTRODUCTION

We present results from the European Commission supported PrestoPrime project and the UK Technology Strategy Board supported AVATAR-m project on analysis and comparison of digital preservation strategies, for example file format migration and the use of different storage models including HDD, data tapes, long-lived media, and encoding schemes with high resilience to data corruption. We start by using a risk assessment methodology (DRAMBORA and OCTAVE Allegro) to identify the origins and impact of various threats to digital AV content from the use of IT systems. We then consider the interplay between cost, risk and loss for audiovisual content when held in IT systems, including the various techniques that can be applied to achieve long term data integrity. Finally, we look at new approaches to how files can be safely stored on imperfect storage systems, including an example of how Dirac encoding can be optimised against data corruption.

BACKGROUND

Digital storage media continues to show an inexorable year-on-year increase in capacity. Hard drives have doubled in capacity every 18 months for the last 30 years (1) and the LTO data tape roadmap as how been extended to 8 generations (2). In another 30 years, an Exabyte (10^{18} bytes) of data will fit on a single storage device, which equates to approximately 1 million hours of uncompressed 1080p HD video or the DPX images for 1 million hours of a film scanned at 2k. The attractiveness of IT storage for archiving large volumes of audiovisual content is obvious. Thankfully, this increase in digital media capacity does not come at an increase in cost (3). It also comes with increased rate at which files can be accessed and transferred, which now allows the archive to be more central in the production, post-production and distribution process(4). As the industry goes ‘tapeless’ and the archive becomes much more central and embedded, archive technology and IT storage and network technology all start to blend together. This satisfies the need for easier and faster access to archive content in both professional and public access scenarios. The benefits apply equally to preservation of existing AV content, for example the BBC D3 project (5), which is removing the need to use a traditional cycle of migrating large numbers of discrete items on specialised carriers from one AV format to another.

Overall, IT based systems, including storage technology, promise lower costs, easier access, and reduced preservation effort, e.g. no more migrations of 'tapes on shelves'. Benefits also include improved archiving, for example capturing and preserving content much earlier in its lifecycle before generation losses occur (6) and capturing essential technical and descriptive metadata at point of creation. But how safe are these IT systems and technologies? What guarantee is there that what goes in today can be retrieved in 50 years time? And if you can get it back out, how closely will it match the original, including the 'bits' but also its faithfulness to the original image or sound?

RISK ASSESSMENT

Risk management is a cyclic activity(7) of assessing and dealing with risk, including the selection and application of one or more treatments. Risk management as a methodology is ideally suited to assessing 'whether IT systems are safe' in the context long-term storage and access of AV assets. Not surprisingly, application of risk management techniques is widespread in critical applications, e.g. information security (8). In the digital preservation domain, the CCSDS (producers of OAIS) are currently combining the efforts of TRAC(9), DRAMBORA(10), Nestor (11) and ISO/IEC 27001:2005 and to ISO standardise the results in the same way as the OAIS Reference Model (12) In PrestoPRIME, we have combined DRAMBORA with OCTAVE (13) for assessment of the threats to data integrity and authenticity from ICT storage technology in audiovisual archiving. The result is a detailed analysis of risks to audiovisual files from the use of IT systems, including origins, assets affected, impact, and suggested mitigation techniques. Some examples of the risks considered are shown in Table 1 with full details in (14).

DRAMBORA Risk ID	Title	Example
R30	Hardware Failure	A storage system corrupts files (bit rot) or loses data due to component failures (e.g. hard drives).
R31	Software Failure	A software upgrade to the system looses or corrupts the index used to locate files.
R32	Systems fail to meet archive needs	The system can't cope with the data volumes and the backups fail.
R33	Obsolescence of hardware or software	A manufacturer stops support for a tape drive, insufficient head life left in existing drives owned by the archive to allow migration
R34	Media degradation or obsolescence	The BluRay optical discs used to store XDCAM files develop data loss.
R35-R38	Security	Insufficient security measures allow unauthorised access that results undetected modification of files.
R39	Disasters	All content is co-located on small-footprint storage systems (e.g. tape robot) that are vulnerable to large-scale loss in a fire or flood.
R40	Accidental System Disruption	An operator accidentally deletes one or more files.
R55, 56, 59	Loss of integrity or authenticity	There is no audit trail for the changes made to content, which mean preservation actions are not taken or are inappropriate.
R60	Unsuitable backups	The backup tapes can't be read.
R61	Inconsistent copies	There are two copies of the content but they are different due to corruption of one, but which one can't be identified.
R64, R69	Content Identifiers	The identifier used to locate a particular file in the system is lost or corrupted.

Table 1 Example risks to AV data from use of IT systems

Risks can be classified into four main areas.

- **Risks of loss of data authenticity and integrity**. These risks are mostly concerned with the loss of ability to track and record the origins of data and then everything that is done to data during digital preservation. Without this provenance trail, there is the risk that changes to integrity or authenticity happen but go unnoticed.
- **Risks of data destruction or degradation**. These risks are concerned with the loss or corruption of data, for example from imperfect storage technology, deliberate or accidental damage, or loss of access to data due to technical obsolescence or which are equally important for archives, then a further set of risks arise
- **Risks to data through loss of services**. If there is a loss or interruption to the services or processes involved in preservation or access to digital content, then this has the potential to put the content itself at risk of loss. For example, this might be the loss of a service that routinely checks and maintains data integrity in a storage system.
- **Risks to loss of data integrity through mismatch of expectations**. If preservation is provided as a service, e.g. within an organisation or by a third-party, then there the potential for a mismatch in expectations or understanding between the providers of the service and the community for which the services are being provided. If change is too rapid, or not communicated properly, then data can be put at risk. For example, the required level of data integrity might not be properly defined, or the sudden need for higher levels of integrity might be beyond the capabilities of current systems.

COST OF RISK OF LOSS

For each of the risks it is possible to reduce or mitigate the risk, but at a cost. The issue is establishing an acceptable balance between increased cost and lowered risk of content loss. This is not simple and the outcome will vary over time requiring constant review. For example, the use video compression means less storage space, which in turn means more copies can be held for the same total cost and a consequent increase in safety. However, each copy is more sensitive to data corruption (18) and compressed formats typically become obsolete faster than uncompressed formats and hence require file format migration on a more regular basis. This adds new costs and risks. The total cost of storage is falling rapidly, so the point at which it becomes more cost effective to store uncompressed is a moveable feast. The issue now becomes one of considering not only risks but the long-term trends for the cost of reducing these risks – for example trends for storage, the longevity of file-formats, and the safety of data in IT systems. This approach is shown in Figure 1. The objective is to convert an archive's needs (how much content it has, how long to keep it, how safe it needs to be, and who needs to be able to access it and how easily) into a preservation plan (what to do, when to do it, what the consequences will be – including accessibility or potential loss of content). Combining preservation modelling (e.g. file level preservation approaches) with storage modelling (bit level preservation approaches) allows the

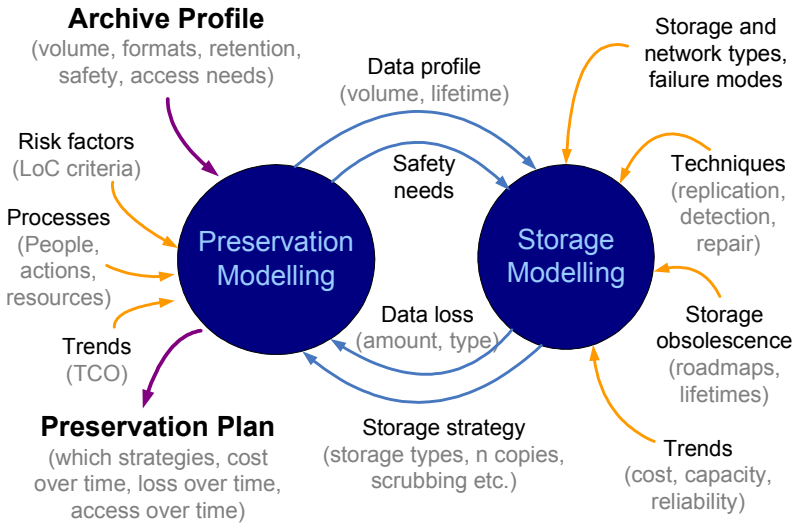

Figure 1 Cost of Risk of Loss assessment and planning

interplay between these two to be considered (e.g. choice of file format impacts on storage required, and formats need evaluating for their sensitivity to data corruption in storage). Calculations can be done on when to make transitions, e.g. from data tape to hard disk, from compressed to uncompressed file formats, or even from in-house to remote archive hosting. This depends on budgets, content volumes, retention schedules, frequency and type of access, content value, maintaining in-house skills, and the IT technology used.

Figure 2 shows an example of this approach for video file format migration. Many long-term sustainability issues are associated with AV file formats, especially modern compressed video formats used in production and post-production. These include tool support, adoption, vendor lock-in, transparency, patents and open standardisation (15). The cost of storage is currently prohibitive for high bit rate material (e.g. one hour of uncompressed HD or 2k film is approx 1TB with a TCO of $1000 per year), but this is falling fast. Whilst compressed formats (and their associated risks) are often considered the only viable option today, the use of uncompressed formats rapidly becomes viable, e.g. within 5 years. The aim is a lifecycle where compression exists only once at the start, if at all, (i.e. no transcoding) and migration to uncompressed happens as soon as possible when lower storage costs are viable.

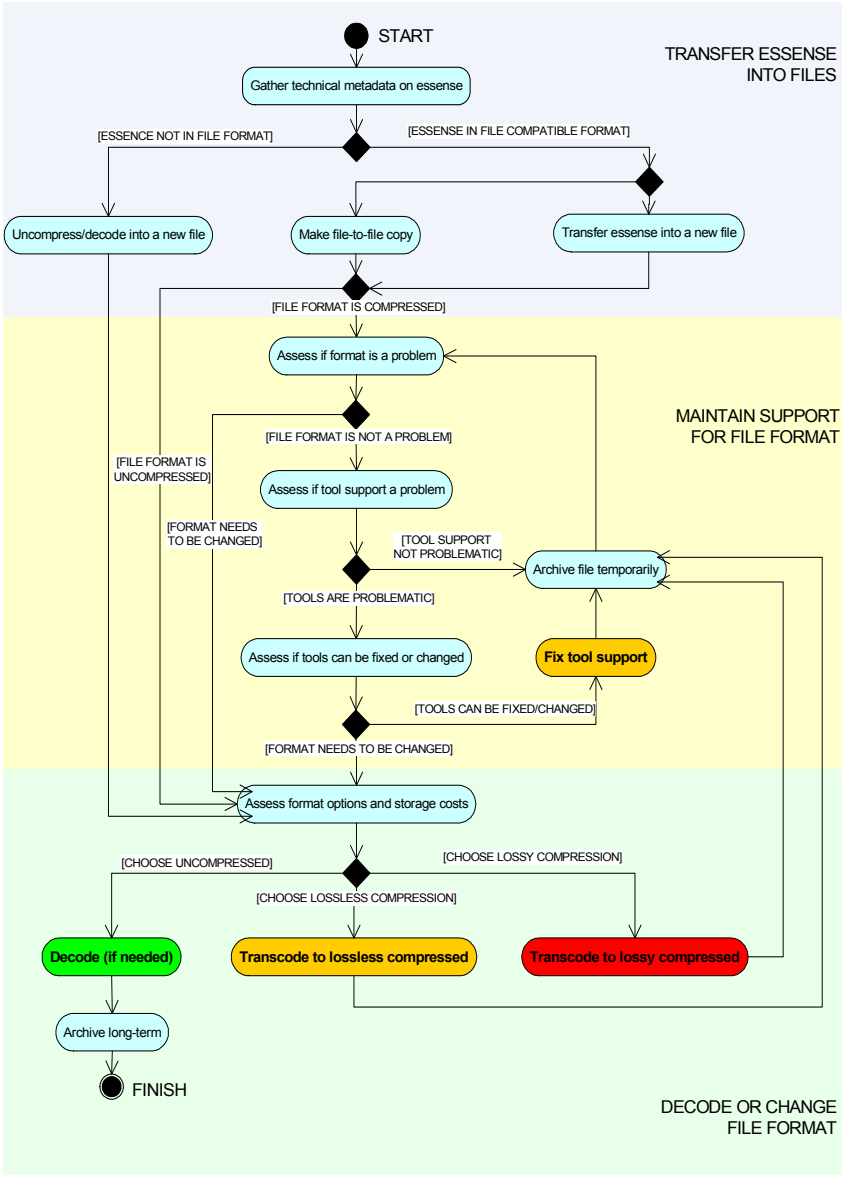

Figure 2 File format migration model

THE RISKS AND COSTS OF USING IT STORAGE TECHNOLOGY FOR AV CONTENT

The major risks to AV content from IT storage technology are two fold. Firstly, short lifetimes require frequent migration to avoid loss from technical obsolescence, for example limited backwards compatibility of LTO data tape drives along with rapid progression through the LTO roadmap means the effective lifetime of a data tape is approximately 6 years. For hard disk drive (HDD) systems, the lifetime is less: 5 years being a typical service life for an individual drive. Secondly, data corruption can take place in all types of IT storage and at all levels, including in systems explicitly designed to prevent it, for example RAID arrays of HDD. Most worryingly, this corruption can be silent and permanent, although it should be said that corruption levels are actually remarkably low,

which is testament to the levels of engineering in these technologies. For example, a modern hard drive has a Bit Error Rate of 1 in 10^{14} with data tape being lower at 1 in 10^{17} - although storage systems with multiple drives or tapes will add further errors that mean the total probability of corruption is often higher than this in practice - see (14) (17) (19) for details. The point here is that whilst error rates may be low, the improvement in error rates is not keeping pace with increased capacity or with the total number of bits in an AV file, e.g. 10^{13} for an hour of uncompressed HD video. The probability of a corruption inside a large AV asset from IT storage is no longer insignificant. The impact of this corruption is also amplified if the file is compressed, e.g. studies show (18) that a single byte corrupted in a JPEG2000 image (lossless or lossy) can result in 30% or more of the decoded pixels being affected and in many cases causing major visual artefacts across the whole image.

It is well known that the TCO of IT storage is much higher than the HDD or data tapes within it, with a factor of 10 being typical when power, space, cooling, people and maintenance are included (16). This TCO falls year-on-year, for both for in-house systems or outsourced services, halving every 2-3 years on average (16). The total lifetime cost of storage can be estimated as a multiplier of today's raw media cost e.g. x10 for the annual TCO and x4 again for the lifetime TCO. The challenge is the large and upfront nature of this cost. The temptation (or necessity) is that compression will save costs – but this adds risks due to format sustainability, increased susceptibility to corruption, and the need to migrate. Herein is the dilemma – what is the best approach, including the alternatives such as long-lived or more reliable technologies, or simply making more copies?

COMPARING AND COMBINING STRATEGIES FOR DATA PRESERVATION

There are many approaches to long term preservation of digital audiovisual content. Each one has associated costs and risks as well as delivering differing degrees of content accessibility. No single technique provides a complete solution. Many archives face the challenge of how to compare, assess and combine the options in a consistent way.

Figure 3 presents a model for analysing preservation strategies for data safety. With reference to the diagram, the bedrock of data safety is to keep multiple copies of content (green circle), using different technologies and in different locations, and ideally operated by different people. This guards against major risks, e.g. disaster recovery, but also unanticipated problems with individual technologies and processes – i.e. it ensures eggs are not 'all in one basket' at any level. For each copy, there is the need to regularly migrate each component of the technology stack (hardware, operating system, management software, formats etc.). There is always the chance that one of the copies is damaged or lost due to some form of failure in the system (orange circle). But only after this problem is detected (yellow circle) can any action can be taken, e.g. to repair or replace the damaged or lost copy. If at any time something happens to the second copy, then there is a risk that both copies are permanently lost or damaged (red) – i.e. content is lost. The rate at which transitions happen between the states dictates how long content is at risk of this loss. Every transition has a cost and hence considering the model as a whole allows the total cost and total risk to be assessed and individual strategies compared as shown in Table 2.

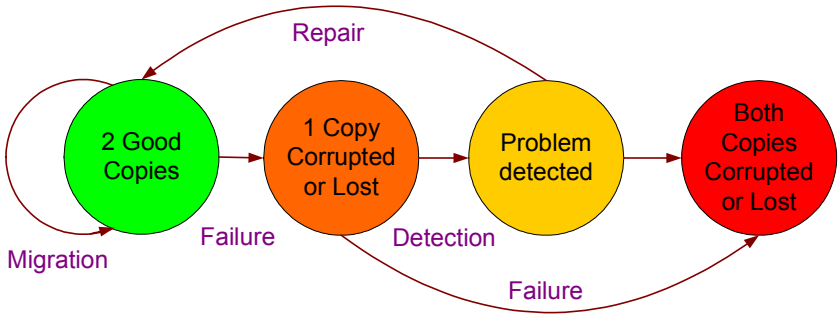

Figure 3 Model of preserving data integrity

Strategy	Example	Migration	Failure	Detection	Repair	Access to content	Notes
Very long lived media	Printing digital bits onto polyester film stock	Infrequent if at all, e.g. film lifetime >200 years	Depends on storage conditions, but very unlikely if good practice followed.	Inspection or spot tests. Hard to automate, i.e. high labour cost	Reprinting in whole or in part. Very expensive	Relatively difficult. Expensive. Latency is measured in days or more. Needs a film scanner	Possibly the only option if there is a risk that 'active' preservation can't be sustained
Reliable media	Data tape, e.g. LTO5	Frequent, e.g. every 6 years or less for LTO tape due to limited backwards compatibility of new drives with old media	Very low bit error rates. Failure rates typ. 0.1-1% of tapes. Problems are often in drives not tapes.	Only need to check integrity on access or during migration	Replace damaged tapes or drives. Drives are expensive and have limited life.	Latency can be high, e.g. tapes on shelves, but data rates good. Need multiple drives for concurrent access.	Other types of reliable media, e.g. magneto optical disks bring other risks, e.g. lock-in to vendors who can go bust.
Many copies	2 online copies on HDD and 2 backup copies on data tape	Frequent, but depends on technology used for copies	The number of individual failures will go up as number of copies goes up	Reduced need to check copies due to increased redundancy	Can repair less often, e.g. only after certain number of copies are lost	More copies can mean easier access, inc. sharing of load for multiple users	Number of copies typically limited by prohibitive costs for video or film
Resilient encoding	Adapted Dirac or JPEG2000 encoding, uncompressed	Format migration for uncompressed is infrequent, e.g. 30 years. Shorter for compressed formats e.g. dirac or JPEG2000	Some data corruption can occur without loss of usability of content, e.g. impact is not visually significant or is correctable.	Need to detect less often due to increased resiliency to corruption.	Repair built in, or 'graceful degradation' means quality is still acceptable and repair not necessary.	Depends on availability of decoders, but not a problem for established formats e.g. JPEG or uncompressed	Virtually all compressed image, audio and video encodings act as huge 'amplifiers' to data corruption.
Concealment	Digital Video Tape	Obsolescence times are short, e.g. 5-10 years	Failures, e.g. read errors, are detected and repaired or concealed automatically by the player, e.g. DV deck.			Hard to automate. AV equipment.	Equivalent in the IT world is digital restoration tools.
Check often, fix quickly	Hard drive storage	Frequent, e.g. every 5 years or less.	Relatively frequent, can be silent and unrecoverable	Proactive checking of file integrity, e.g. using checksums	Replace damaged copies. Can need large data transfers, e.g. TB files to fix only a few bits of corruption	Low latency, high bandwidth. Random access to parts of files, e.g. 'partial restore'. Easy to support many users.	Latent errors can occur at all levels of the storage stack, including in parts designed to protect data, e.g. RAID

Table 2 Comparison of data storage strategies

RESILIENT ENCODING

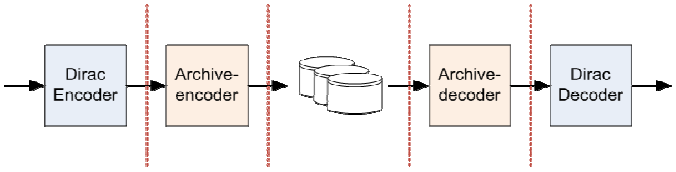

Figure 4 Extended Dirac encoding/decoding scheme to add resilience against data corruption

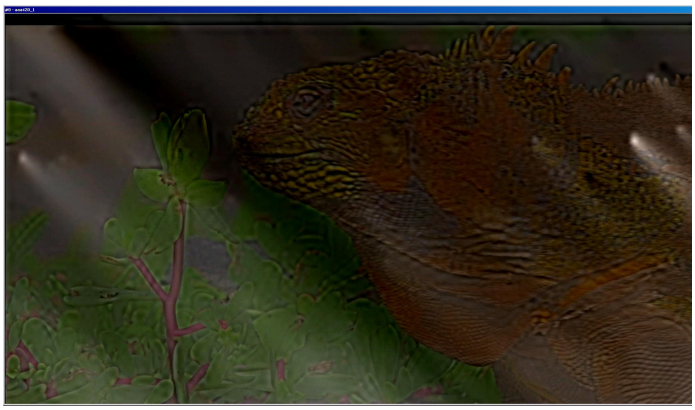

Figure 5 Effect of data corruption on the Dirac DC band (reproduced with permission of BBC)

Figure 6 Effect of data corruption on the low freq sub band (reproduced with permission of BBC)

Figure 7 Effect of data corruption on the high freq sub band (reproduced with permission of BBC)

One approach that is being investigated by the BBC to counter the effects of data corruption from IT storage technology is to use encoding schemes that allow more protection to be given to the parts of a video file that are most sensitive to corruption, for example the header information or the lower frequency coefficients in wavelet based video compression schemes such Dirac or JPEG2000. An extra encoding stage is performed just before the content goes to storage (Figure 4) and then this is reversed by a decoding stage on retrieval. In this way, the original video encoding is not changed, only the way that the content is written to/from one or more storage systems. For example, a single Dirac video file can be split into component files (sub files) by grouping sub bands of frequency coefficients. The header metadata is replicated in each sub file to add redundancy at trivial extra storage cost. Each sub file is then stored according to its sensitivity to corruption, e.g. on different storage technologies or with different levels of replication. If corruption occurs in the lower frequency sub bands then ability to use the content is completely lost (Figure 5 and Figure 6), therefore the corresponding sub files are given the most protection. If corruption occurs in the higher frequency sub bands, then loss may be tolerable (Figure 7). These sub files can be given less protection. In this way, maximum protection (highest cost) is given to the parts of the file where the potential effect of loss is most significant. Work is underway to apply a similar approach to JPEG2000, with the benefits of being able to protect against much larger blocks of corruption than can be accommodated by JPEG2000 inbuilt correction scheme (e.g. JPEG2000 over wireless ISO/IEC 15444-11:2007).

CONCLUSIONS

Many risks arise when IT storage technology and systems are used for long-term AV data integrity and usability. IT technology, despite its imperfections, can achieve much higher levels of safety than previously possible when archiving AV material as 'items on shelves'. Whatever the level of safety needed, and the measures used to achieve it, the issue is how much it costs, what are the risks of loss of content, and what is the benefit of incurring more cost to further reduce these risks. We have shown a structured approach based on risk assessment, cost modelling and new ways to achieve preservation of AV files.

ACKNOWLEDGEMENTS

AVATAR-m is a UK TSB supported collaborative R&D project involving the BBC, Ovation Data Services, Xyratex and IT Innovation. For more information please contact avatar-m@it-innovation.soton.ac.uk or see www.avatar-m.org. PrestoPrime is a EC supported 7th Framework Programme ICT project (FP7-231161) coordinated by INA (Institut national de l'audiovisuel) in France. Partners include BBC, RAI, ORF, B&G and others. For information contact prestoprime@it-innovation.soton.ac.uk or see www.prestoprime.org

REFERENCES

(1) http://en.wikipedia.org/wiki/Moores_law

(2) http://www.ultrium.com/index.html

(3) http://www.mattscomputertrends.com/harddrives.html

(4) Thomas M. Coughlin, 2009, Digital Storage for Professional Media and Entertainment. http://www.tomcoughlin.com/

(5) Stuart Cunningham and Philip de Nier, 2007, File-based Production: Making It Work In Practice. http://downloads.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP155.pdf

(6) Are Olafsen, 2010, Compressing 3G Video, in proceedings of NAB 2010.

(7) www.theirm.org/publications/documents/Risk_Management_Standard_030820.pdf

(8) CERT: http://www.cert.org/work/organizational_security.html

(9) TRAC: http://www.crl.edu/PDF/trac.pdf

(10) DRAMBORA: http://www.repositoryaudit.eu/

(11) Nestor: http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf

(12) OAIS Blue Book: http://public.ccsds.org/publications/archive/650x0b1.pdf

(13) OCTAVE: http://www.cert.org/octave/

(14) Matthew Addis et al, 2010. Threats to data integrity from use of large-scale data management environments PrestoPRIME Deliverable ID3.2.1 http://www.prestoprime.eu/

(15) http://www.digitalpreservation.gov/formats/sustain/sustain.shtml

(16) Matthew Addis et al, 2010. Audiovisual preservation strategies, data models and value-chains PrestoPRIME Deliverable D2.2.1 http://www.prestoprime.eu/

(17) http://blog.dshr.org/2008/03/more-bad-news-on-storage-reliability.html

(18) Heydegger, V (2009) Just One Bit in a Million: On the Effects of Data Corruption in Files. Research and Advanced Technology for Digital Libraries, ECDL 2009, LNCS 5714

(19) M. Addis, R. Lowe, L. Middleton, N. Salvo. (2009) Reliable Audiovisual Archiving Using Unreliable Storage Technology and Services. In proceedings of IBC 2009.